THE ROYAL SOCIETY

Published online 31 October 2002

Credible threats and promises

John M. McNamara[1*] **and Alasdair I. Houston**[1,2]

[1]*Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK*
[2]*School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK* (*a.i.houston@bristol.ac.uk*)

We consider various implications of information about the other player in two-player evolutionary games. A simple model of desertion shows that information about the partner's behaviour can be disadvantageous, and highlights the idea of credible threats. We then discuss the general issue of whether the partner can convince the focal player that it will behave in a specific way, i.e. whether the focal player can make credible threats or promises. We show that when desertion decisions depend on reserves, a player can manipulate its reserves so as to create a credible threat of desertion. We then extend previous work on the evolution of trust and commitment, discussing conditions under which it is advantageous to assume that a partner will behave in a certain way even though it is not in its best interest.

Keywords: commitment; desertion; cooperation; threats and promises; trust; information about intentions

1. INTRODUCTION

It is well established that natural selection can result in one animal giving reliable information about itself to another animal (reviewed in Johnstone 1997, 1998; Bradbury & Vehrencamp 1998). Many models of this process involve an animal giving an honest signal about its quality. At the evolutionarily stable signalling strategy, high-quality individuals do better than low-quality individuals. In this paper, we are concerned with a different aspect of information transfer. We ask whether animals might be selected to give accurate information about their future behaviour, i.e. about what we might think of as their intentions (e.g. Maynard Smith 1979; Krebs & Dawkins 1984). We refer to this disclosure of information as making a threat or a promise. In some contexts the threat or promise is to perform an action that is in the animal's best interests (see, for example, Enquist 1985; Hurd & Enquist 1998), so that the threat or promise is credible. The possibility of an animal giving reliable information about its future behaviour becomes doubtful when the intended behaviour is not in the animal's best interests. The animal could make the information reliable if it could commit itself to the particular course of action so that it cannot deviate from it when it becomes advantageous to do so. This is the 'commitment problem' (see Schelling 1960; Elster 1984, 2000; Frank 1988; Samuelson 2001). We consider two examples that illustrate two ways in which this problem can be solved. In the first example, an animal manipulates its state in such a way as to restrict its options. In the second example, individuals differ in their pattern of behaviour, and some idea of an animal's type can be obtained from its appearance or behaviour. One type (called 'trustworthy') commits itself to a given action (cooperation if trusted) despite the fact that this is not its best choice. We consider a single interaction between two individuals. In this case there can be no effect of the reputation established in previous interactions, and there is no advantage to be gained from future interactions. Frank (1987, 1988, 1989) and Guth & Kliemt (2000) have shown that trustworthy individuals can persist in such circumstances. We analyse a modified version of their model, and then extend their analysis to the case where individuals choose a costly signal of trustworthiness.

2. EXAMPLE 1: A PARENTAL CARE GAME

Consider a male and female that have just produced young. These parents play a game against one another in which each has to decide whether to care for the young or to desert them. Survival of the young is enhanced by parental care, but each parent can also enhance its own reproductive success if it deserts, mates again and produces extra young with its new mate. For example, suppose that the two parents are equally good at care. If no parents care then all the young die. If one parent cares then four young survive. If both parents care then six young survive. We assume that a female that deserts produces three extra surviving young from other matings. The male only produces one extra young if he deserts. This asymmetry might, for example, be because the sex ratio in the population is male biased, so that it is harder for a male to find a new mate. We take the pay-off to a parent to be the total number of surviving offspring of that parent. With the above assumptions, the pay-offs are given by tables 1 and 2. From table 1 it can be seen that if the male knows the female's decision, then he should care regardless of whether she cares or deserts. By contrast, from table 2 it can be seen that if the female knows the male's decision, then she should care if the male deserts and desert if he cares.

[*] Author for correspondence (john.mcnamara@bristol.ac.uk).

One contribution of 12 to a Theme Issue 'Information and adaptive behaviour'.

Table 1. Pay-offs to the male.

	female cares	female deserts
male cares	6	4
male deserts	5	1

Table 2. Pay-offs to the female.

	male cares	male deserts
female cares	6	4
female deserts	7	3

We need to distinguish between the action that a parent chooses ('care' or 'desert') and the strategy of that parent. The strategy is a rule specifying how the action is chosen. What strategies are possible depends on the circumstances; that is it depends on the details of the process by which decisions are reached. Natural selection acts on strategies. A pair of strategies for the parents are in Nash equilibrium if the strategy of the male is a best response to that of the female and vice versa. If, in addition, each strategy is the unique best response to the other, the pair of strategies is evolutionarily stable in the sense defined by Maynard Smith (1982) and proved by Selten (1980). To illustrate these concepts, we contrast the scenarios (a) and (b) below.

(a) *Simultaneous choice*

Suppose that each parent chooses its action before it knows the action chosen by its partner. Neither can alter its decision once the choice of the partner subsequently becomes known. This scenario is usually described as one of simultaneous choice, but it is the lack of information at the time of choice, rather than the timing of decisions, that is crucial to the game. Here, a strategy simply specifies the probability that the parent cares (and hence also the probability that the parent deserts). Thus, for example, the strategy 'always care' specifies that the parent cares with probability 1.

If the male knows the female's choice then he should care regardless of that choice. Thus, if the male does not know the female's choice he should care. Given that the male cares the female does best to desert. Thus, the male strategy 'always care' and the female strategy 'always desert' are unique best responses to one another, and the only evolutionarily stable pattern of care is uni-parental care by the male.

(b) *Male chooses first*

Suppose that the male decides first. The female then makes her choice on the basis of the male's decision. In this scenario there is an informational asymmetry: the female knows the male's decision when she makes her choice; whereas the male does not know the female's decision when he makes his choice. As before, the information structure, rather than the timing of decisions, is crucial to the game. A strategy for the male again specifies the probability that he cares. A strategy for the female now specifies the probability that she cares when the male has chosen to care and the probability that she cares when the male has chosen to desert.

Consider the female strategy specified by the rule: if the male cares then always desert, if the male deserts then always care. We refer to this strategy as 'sensible'. Suppose that the female uses this strategy and consider the best response of the male. If the male cares then the female deserts. The result is uni-parental care by the male, and he obtains a pay-off of 4 (table 1). Suppose instead that the male deserts. The female then cares, resulting in a pay-off of 5 to the male (table 1). Thus, the male's best response to the female strategy 'sensible' is the strategy 'always desert'. Essentially, it is best for the male to desert to prevent the female from deserting. Conversely, if the male uses 'always desert' then any strategy that specifies 'care' when the male deserts is a best response for the female. The strategy 'sensible' has this property. Thus, the male and female strategies 'always desert' and 'sensible' are best responses to each other and are in Nash equilibrium. In a population at this equilibrium there is uni-parental care by the female.

Suppose that all males in a population use the strategy 'always desert' and all females use 'sensible'. Is this population evolutionarily stable? To analyse this let 'care regardless' be the female strategy specifying that the female always cares regardless of the action chosen by the male. This strategy is also a best response to the male strategy 'always desert'. If a mutation gives rise to some females using 'care regardless', then these females do equally well as those using 'sensible', even when mutant numbers rise in the population. This occurs because 'sensible' and 'care regardless' only differ as strategies in their specification of what to do when the male cares, but as males never care the difference is never manifest. Because mutants are not selected against, according to the original definition given by Maynard Smith (1982) the population is not evolutionarily stable (Selten 1983). The situation alters, however, if males make 'mistakes'. Suppose that all males in the population use the strategy 'always desert', but an occasional error is made resulting in some males choosing to care. Females using the strategy 'sensible' will then do better than those using 'care regardless', because when a male does care it is best for the female to desert. Thus, the occurrence of occasional errors stabilizes the population. Selten (1983) refers to an equilibrium that is stable under occasional error as a limit ESS.

In a population at the above equilibrium, for each possible choice of action of a male the female makes the best choice in the circumstance. (It is also true that the single action chosen by the male is the best in the circumstance.) As such the solution of the game is an example of a sub-game perfect equilibrium (e.g. Fudenberg & Tirole 1991). Selten (1983) has argued that evolutionary game theory needs to take account of the fact that organisms will make occasional errors. The resulting limit ESSs will then be sub-game perfect (Selten 1983). We agree that errors are important and that for many scenarios we should be seeking sub-game perfect equilibria. We do, however, have a proviso. Sub-game perfection demands that individuals are able to take the best action in any circumstances. We argue below that organisms may lack the flexibility required by sub-game perfection. The behaviour of an organism is likely to be given by a rule that does not have

a unique response for every situation. An organism following such a rule will be inflexible, so that past behaviour may indicate that future behaviour will not necessarily be in the animal's best interests. It is then not reasonable to expect sub-game perfection, and as a result certain threats and promises become credible.

The contrast between the two scenarios that we have considered highlights two points that apply to evolutionary games in general. In the simultaneous choice game the evolutionarily stable outcome is male-only care. By contrast, in the game in which the male chooses first there is female-only care. Thus, the decision process by which the eventual outcomes are chosen is crucial to the evolutionarily stable outcomes of the game (e.g. Hurd & Enquist 1998; Houston & McNamara 1999; McNamara *et al.* 1999).

The second general point is illustrated by comparing the pay-offs at equilibrium in the two games. Male-only care in the simultaneous choice game results in a pay-off of 4 to a male and 7 to a female (tables 1 and 2). Female-only care in the male chooses first game results in a pay-off of 5 to a male and 4 to a female (tables 1 and 2). The only difference between these games is that the female is ignorant of the male's choice when she makes her own choice in the first game, whereas she knows the male's choice in the second game. The extra information available to the female has therefore put her at a disadvantage and has been beneficial to the male. This illustrates the general point that a contestant in a game may be able to gain an advantage if he is able to reliably signal his intention to his opponent. It also shows that the nature of the decision process can determine an individual's pay-off. This indicates that there will be selection pressure acting on the form of the process.

The comparison between the two scenarios is analogous to an effect of omniscience described by Brams (1983). Brams is concerned with various games played by an ordinary human player, P, and a superior being, SB. In one of the cases that Brams considers, the SB is omniscient, i.e. it can predict the choice that P will make. If P knows that SB has this ability, then this may increase P's pay-off and decrease SB's pay-off (Brams 1983, ch. 4). Brams illustrates this effect in the context of the game known as 'chicken', but the effect also occurs in the desertion game. Omniscience transforms a simultaneous choice in which each player is ignorant of the other's choice into a choice procedure in which one player knows the other's choice. This is equivalent to P deciding first and SB deciding second. As in the desertion game, the player deciding second does worse than when choice is simultaneous.

(c) *A threat that is not credible*

Consider again the parental effort game in which the male chooses first. We have seen that this game has a Nash equilibrium at which the male adopts 'always desert' and the female adopts 'sensible'. At this equilibrium, there is care by the female alone. We now consider an equilibrium at which there is care by the male alone. Let 'desert regardless' be the female strategy that specifies desertion regardless of the male action. Then it can be seen that the male strategy 'always care' and the female strategy 'desert regardless' are in Nash equilibrium. Here, the female's threat to desert even if the male deserts forces him to care. At the Nash equilibrium males always care and hence females never need to carry out their threat. But is this threat strategy liable to evolve?

'Desert regardless' specifies that a female should desert if the male has already deserted. But, given the male has deserted, the best action in this circumstance is to care. Thus, the Nash equilibrium is not sub-game perfect and we can mirror the argument of the last section as follows. Consider a population in which all the males follow 'always care' and all females follow 'desert regardless'. Then females following the mutant strategy 'sensible' will do equally well as the resident females even if mutant numbers increase. Furthermore, if males make occasional errors by deserting, then 'sensible' will do strictly better than 'desert regardless'. Thus, the population is not evolutionarily stable. In this particular example it is not necessary to assume that male 'errors' are due to sub-optimal behaviour. If we take male desertion to mean that the male fails to return to the young, then this could be the result of the male being killed by a predator while foraging away from the young.

This example illustrates that threats in which an individual promises to do something that is not in its self interest are not credible. That is, we would not expect such empty threats to be used to settle contests in nature, at least if behaviour is completely flexible in the sense described above.

(d) *State-dependent desertion: an example of a credible threat*

Barta *et al.* (2002) model the care and desertion decisions of members of a population of birds during a breeding season. In the most basic form of the model, birds do not expend energy and hence do not need to feed. Individuals that are single search for a mate. Individuals that have found a mate produce young and then decide whether to care for the young or desert. The male decides first. The female then makes her choice knowing the male's choice. Individuals that desert are again single and immediately start to search for a new mate. Individuals that care only become single after the young have reached independence. Thus, care wastes time that could be spent finding and mating with a new partner. The predictions of the model depend on parameter values. In one region of parameter space Barta *et al.* find that at evolutionary stability young are cared for by both parents at certain times in the breeding season and by the female alone at other times. When female-only care occurs it does so because the male, who chooses first, deserts to prevent the female from deserting, as in the example based on tables 1 and 2.

Barta *et al.* (2002) compare the predictions of their most basic model with predictions of a modified model in which there are energy requirements and two types of decision. As before, individuals with young decide whether to care, but now single birds must decide when to feed and when to search for a mate. The energy reserves of a bird increase when it feeds, but decrease during mate search, production of the young and care of the young. If reserves reach zero then the bird dies of starvation. This means that when a bird mates it has to have sufficient reserves to survive until the young are produced. If it will also care for the young, reserves need to be higher still at mating.

There is no mate choice, so that pairing is random with respect to reserves. However, when parents decide whether or not to care they know both their own reserves and that of their partner. In this version of the model, at evolutionary stability single males feed more than single females. Consequently, within most mated pairs the male has high reserves and the female has reserves that are too low for her to be able to care. Thus, the female has to desert whatever the decision of the male, and even though the male chooses first, he is forced to care. Thus, in contrast to the model without reserves, the predominant pattern of care in this version of the model is male-only care.

In this game females handicap themselves by having low reserves. This then commits them to desert. In other words it creates a credible threat. The threat is credible because it involves sub-game perfect behaviour. In the desertion game with no reserves, the female cannot produce a credible threat of desertion. When the model is modified to include reserves the situation is very different. By reducing her reserves, the female can commit herself to deserting if she is deserted.

3. EXAMPLE 2: TRUST AND COOPERATION

(a) *Abandoning sub-game perfection*

In the example based on desertion, the female was able to generate a credible threat by reducing her reserves to such a level that, if deserted, her best option would be to desert. Thus, the threat is credible because it is the best choice, and so is sub-game perfect. In the current example, we look at a case in which behaviour is not sub-game perfect. Before introducing the example, we outline why sub-game perfect behaviour may not always occur.

If an animal's behaviour is to be sub-game perfect, then it must adopt the optimal decision in any situation that it may encounter. This is likely to require a complex and flexible decision-making procedure. It can be argued that such a procedure is expensive in terms of underlying neurons and imposes far too high a computation load on its user. A more plausible suggestion is that animals use relatively simple rules which perform well under the circumstances that an animal is likely to encounter (McNamara & Houston 1980; Houston *et al.* 1982; McNamara 1996; Houston & McNamara 1999). These rules will result in less flexibility than is required for sub-game perfection. In addition to the idea of simple rules, there is evidence for consistent individual differences in behaviour within a species (e.g. Wilson *et al.* 1994; Heinsohn & Packer 1995; Wilson 1998; Budaev *et al.* 1999). This amounts to saying that individuals have characteristic ways of behaving i.e. have different 'personalities'. If one animal can obtain information on the personality of another animal then it has information on how that animal will behave in future. We now explore some implications of this idea using a simple example with two personality types.

(b) *The basic trust model*

Building on the work of Frank (1987–1989), Guth & Kliemt (2000) present the following two-player game. This game has a role asymmetry, with the player in role 1 choosing first and then the player in role 2 responding. We refer to the player in role i as player i. In the first stage player 1 decides whether to reject or trust player 2. If player 2 is rejected both players receive a pay-off of s and the game ends. If player 1 trusts player 2 then player 2 decides whether to cooperate with player 1 or defect. If player 2 cooperates then both receive a pay-off of r. If player 2 defects then player 1 receives a pay-off of 0 and player 2 receives a pay-off of 1. It is assumed that

$$0 < s < r < 1. \tag{3.1}$$

This is a game of trust and cooperation in which players meet just once. There are thus no effects that occur because games are repeated with the same opponent. It is also assumed that no other population member observes a game, so that there are no reputation effects.

To analyse the game suppose that player 1 trusts player 2. Then player 2 obtains a pay-off of r by cooperating and a pay-off of 1 by defecting. Thus, because $r < 1$, it is best for player 2 to defect. Given that player 2 will defect, player 1 obtains a pay-off of s by rejecting player 2 and a pay-off of 0 by trusting player 2. Thus, because $0 < s$ player 1 does best to reject player 2. Thus, the unique sub-game perfect equilibrium solution of the game is for player 1 to reject player 2 and for player 2 to defect if this player ever has the opportunity to do so. This solution is not an ESS in the strict sense defined by Maynard Smith (1982), as at equilibrium player 2 never has to make a choice and so any strategy by player 2 is a best response to player 1's strategy. The solution is, however, a limit ESS.

(c) *Signalling trustworthiness*

Guth & Kliemt (2000) go on to consider a large population of individuals that pair at random and play the above game. Pair members are assigned to roles at random. Individuals in the population are of two types, labelled trustworthy and untrustworthy. Type is genetically determined and so tends to be inherited by offspring. Type does not affect behaviour when in role 1. Trustworthy individuals cooperate when in role 2. Untrustworthy individuals defect when in role 2. Within a game, before player 1 chooses whether to trust player 2, player 1 can obtain limited information on the type of player 2 by observing this individual. In their analysis Guth & Kliemt (2000) are concerned with the evolutionarily stable proportion of each type in the population. Here, we consider this and other issues using a modification of their model.

Guth & Kliemt (2000) assume that observations can only take two possible values. Here, we modify the model, assuming observations are normally distributed random variables. To be specific, we assume that the value observed by player 1 has a normal distribution with mean -1 and variance σ^2 when player 2 is not trustworthy, and has a normal distribution with mean $+1$ and variance σ^2 when player 2 is trustworthy. At present we assume that individuals of a given type have no control over how they appear. Later we will consider a modification of the model in which the observations are signals chosen by individuals, with higher signals incurring higher cost. In the model of Guth & Kliemt (2000), player 1 has the choice of whether to observe player 2 or not, observations being costly to player 1. Here, we begin by assuming that there is no observation cost to player 1. Later we will consider how observation costs modify our conclusions.

Suppose that the population is composed of a proportion $1 - p$ of untrustworthy individuals and a proportion p of trustworthy individuals. Consider a game played between two randomly selected members of this population. It is assumed that contestants know p. To start with, consider the action of player 1 if this player could not observe player 2. Then, all player 1 would know is that player 2 is trustworthy with probability p. If player 1 rejects player 2 then player 1 obtains a pay-off of s. If player 1 accepts player 2 then player 1's expected pay-off is

$$(1 - p)0 + pr, \tag{3.2}$$

so that the pay-off is pr. Thus, player 1 should trust player 2 if and only if

$$p > \frac{s}{r}\ . \tag{3.3}$$

Now suppose that player 1 does observe player 2, as we are assuming in our model. Then before player 1 has observed player 2 the probability that player 2 is trustworthy is p. Let player 1's observation of player 2 be x. Then the prior probability p can be updated to a Bayesian posterior probability $\pi(p, x)$ that player 2 is trustworthy. By exactly the same reasoning as above, player 1 should trust player 2 if and only if

$$\pi(p, x) > \frac{s}{r}\ . \tag{3.4}$$

The mean observed value of a trustworthy individual is greater than that of an untrustworthy individual. Thus, for given p the probability of trustworthiness $\pi(p, x)$ increases with the observation x. It follows that player 1 should trust player 2 if and only if the observation x exceeds the critical threshold $x_c(p)$ given by $\pi(p, x_c(p)) = s/r$. It is shown in Appendix A that

$$x_c(p) = \frac{1}{2}\log\left(\frac{1 - p}{p}\right) + \frac{1}{2}\log\left(\frac{s}{r - s}\right). \tag{3.5}$$

Figure 1*a* illustrates the dependence of the critical threshold on p. For given observation x, the probability $\pi(p, x)$ that player 2 is trustworthy increases with the proportion of trustworthy individuals p. Thus, the higher p the lower the critical level $x_c(p)$.

The consequences to player 1 of using the critical acceptance threshold $x_c(p)$ are illustrated in figure 1*b*. The probability that player 1 trusts player 2 increases from 0 when $p = 0$ to 1 when $p = 1$. By inequality (3.4), if player 1 trusts player 2 then the probability that player 2 is trustworthy is at least s/r. This probability rises to 1 as p increases. The consequences of player 1's strategy for player 2 are illustrated in figure 1*c*. When p is small, the prior probability that player 2 is trustworthy is small, and player 1 is very choosy. Thus, although the probability of any player 2 being trusted is low, a trustworthy player 2 is much more likely to be trusted than an untrustworthy player 2. Thus, trustworthy individuals do better, and there is selection for the proportion of trustworthy individuals in the population to increase. When p is large, the prior probability of trustworthiness is large, and player 1 is not very choosy. Consequently, untrustworthy individuals are usually trusted. They therefore obtain the maximum pay-off of 1 and do better than trustworthy individuals. There is thus selection for the proportion of untrustworthy individuals to increase. At the evolutionarily stable proportion of trustworthy individuals, both types of individual do equally well.

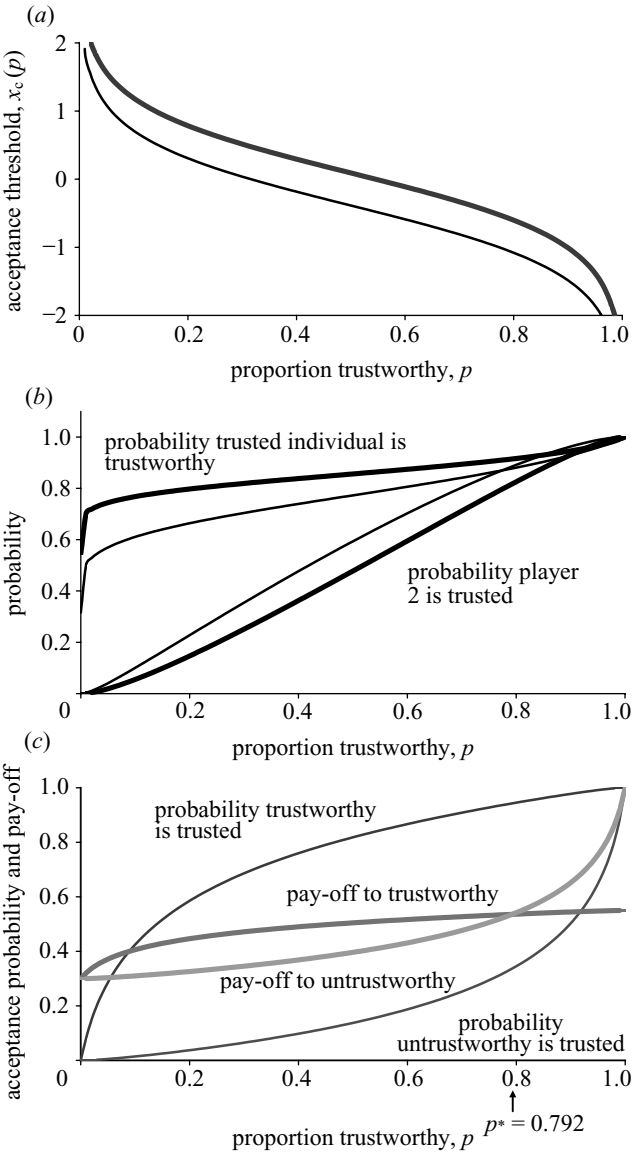

Figure 1. The effect of the proportion of trustworthy individuals on the behaviour of player 1, and the consequences of this behaviour for player 2. (*a*) Acceptance threshold for player 1. (*b*) The probability that player 1 trusts player 2 and the probability that a trusted individual is trustworthy. (*c*) The probability that player 2 is trusted and the pay-off to this player, shown for each type of player 2. At evolutionary stability both types of player 2 do equally well. In the case illustrated the evolutionarily stable proportion of trustworthy individuals is $p^* = 0.792$. Parameters are $s = 0.3$ throughout; in (*a*) and (*b*) bold line, $r = 0.55$; thin line, $r = 0.95$; in (*c*) $r = 0.55$.

We will denote the evolutionarily stable proportion of trustworthy individuals in the population by p^*. The following two general results about evolutionary stability are easily proved (see Appendix B for details). First, because both types do equally well at evolutionary stability, at this equilibrium an untrustworthy individual is $(r - s)/(1 - s)$ times as likely to be trusted as a trustworthy individual; i.e.

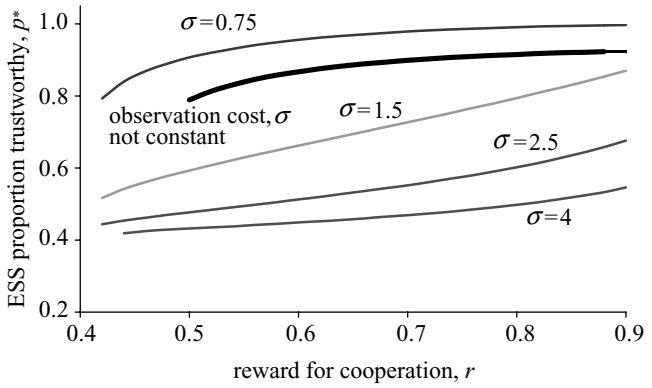

Figure 2. The evolutionarily stable proportion of trustworthy individuals, p^*. This proportion is shown as a function of the reward for cooperation r. The cases where the observation standard deviation σ is fixed are from the model with no observation costs. A case in which σ is determined by the observation cost paid by player 1 is also shown (σ given by equation (C 1) with $c_0 = 0.1$). The parameter $s = 0.4$ in all cases.

$$P(\text{trusted}|\text{untrustworthy}) = \left(\frac{r-s}{1-s}\right)P(\text{trusted}|\text{trustworthy}). \tag{3.6}$$

We also know from inequality (3.4) that the probability that an individual that is trusted is trustworthy is at least s/r. Combining this with equation (3.6) we have

$$p^* > s. \tag{3.7}$$

These results do not depend on the assumption that observations are normally distributed.

Figure 2 illustrates how p^* depends on parameters when observations are normal. As the reward for cooperation, r, increases there are two counteracting effects. Because player 1 becomes less choosy, untrustworthy individuals are more likely to be trusted and hence these individuals do better. But trustworthy individuals do better both because they are also likely to be trusted, and because the pay-off if they are trusted is greater. The consequence of these two effects is that p^* is robust to changes in r. Results are most sensitive to the standard deviation in observation, σ. As σ increases, p^* decreases until it approaches its lower theoretical limit of s (inequality (3.7)).

(d) *Observation costs*

We now modify the basic model to allow player 1 to choose the accuracy of the observation on player 2. Specifically, player 1 chooses the cost c that will be paid for the observation, where $c \geqslant 0$. If no cost is paid then no information about type is obtained. The higher the cost chosen, the lower the variance of the observation σ^2. Thus, the more player 1 pays in cost, the better player 1 can discriminate between the two types of player 2. Details are given in Appendix C.

Typical results are illustrated in figure 3. When p is small it is *a priori* likely that player 2 is untrustworthy. It is therefore not worth paying an observation cost ($c = 0$), and player 2 is rejected without observation. When p is large it is likely that player 2 is trustworthy. It is again not worth paying an observation cost and player 2 is trusted without observation. At intermediate levels of p it is worth

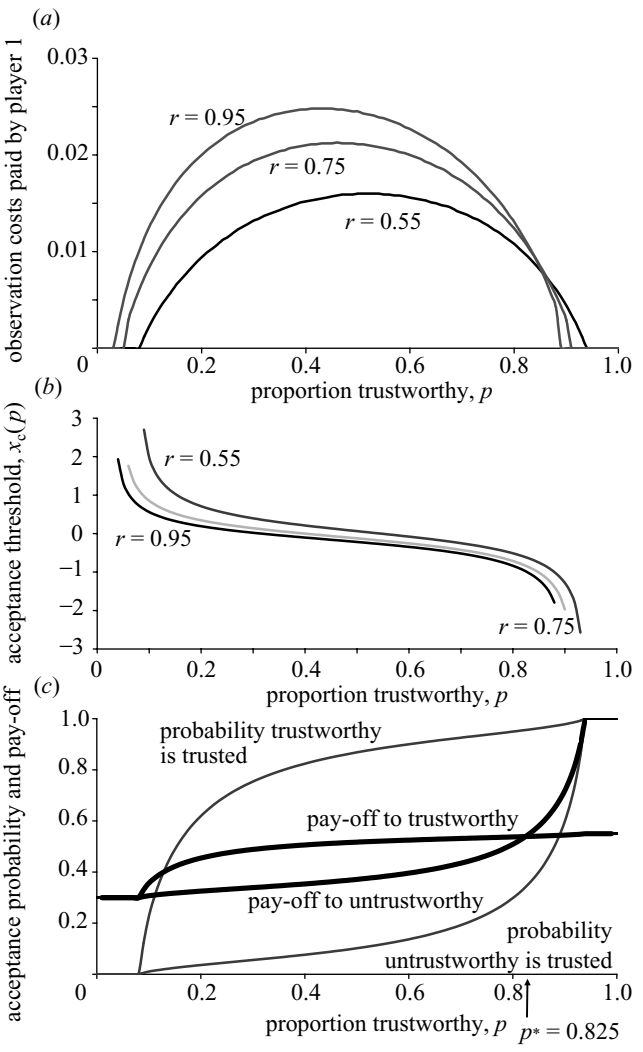

Figure 3. The effect of the proportion of trustworthy individuals for the model in which the observation standard deviation σ is determined by the observation cost paid by player 1. (*a*) The observation cost paid by player 1. (*b*) The acceptance threshold of player 1. (*c*) The probability that player 2 is trusted and the pay-off to this player, shown for each type of player 2. The evolutionarily stable proportion of trustworthy individuals is $p^* = 0.825$. Parameters are $s = 0.3$ throughout; in (*a*) and (*b*) $r = 0.55$, $r = 0.75$, $r = 0.95$ as indicated on figure; in (*c*) $r = 0.55$; σ given by equation (C 1) with $c_0 = 0.1$.

paying for information, with the cost that is worth paying being greatest around $p = 0.5$. The effect of the behaviour of player 1 on the pay-offs to each type of player 2 is illustrated in figure 3*c*. At small p both types are rejected and hence both receive the same pay-off s. Thus, unlike the case with no observation cost, there is a region with neutral selection for low p. Guth & Kliemt (2000) obtain similar results for their model. They argue that player 1 may occasionally make a mistake and accept player 2 in this region. If this occurred then the untrustworthy individuals would do strictly better than trustworthy individuals. There would then be two evolutionarily stable proportions of trustworthy individuals; one with no trustworthy types present and the other with a positive proportion of this type. Between these stable proportions is an invasion barrier with selection against trustworthy types at low frequencies and selection for them at intermediate

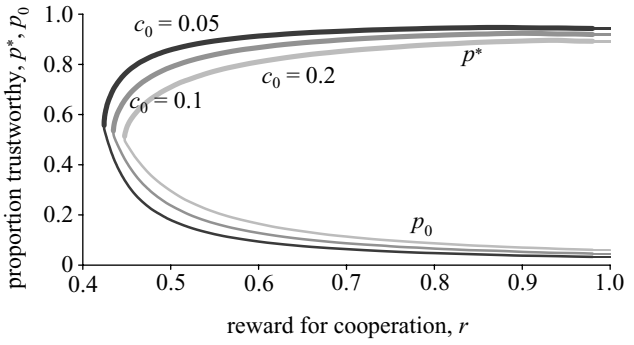

Figure 4. The evolutionarily stable proportion of trustworthy individuals, p^*, and the invasion barrier, p_0, for the model in which σ is determined by the observation cost paid by player 1 (equation (C 1)). The effect of the reward for cooperation r is shown for the three cases $c_0 = 0.05$, $c_0 = 0.1$ and $c_0 = 0.2$. $s = 0.4$ in all cases.

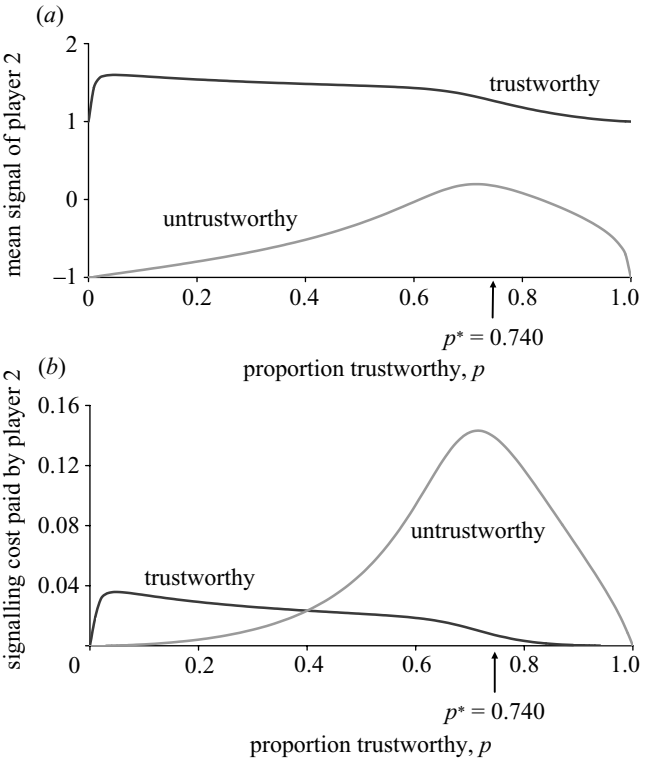

Figure 5. The effect of the proportion of trustworthy individuals in the model in which player 2 can adjust the mean of the signal that it gives. (*a*) The mean signal of each type. (*b*) The cost of the signal in (*a*). The evolutionarily stable proportion of trustworthy individuals is $p^* = 0.740$. Parameters are $s = 0.4$; $r = 0.7$; $\sigma = 1$ $k = 0.1$.

frequencies (with selection against them again at high frequencies). The argument in terms of occasional errors seems reasonable, and we too interpret results in this way.

Figure 4 shows how the invasion barrier and evolutionarily stable proportion of trustworthy individuals depend on the cost structure and the reward for cooperation. For r very close to s it does not pay for player 1 to observe player 2 for any value of p. Acceptance then follows the rule given by criterion (3.3); i.e. when $p < s/r$ player 1 rejects player 2, and when $p > s/r$ player 1 accepts player 2. The only evolutionarily stable proportion of trustworthy individuals is then $p = 0$. When r is not too close to s there are two stable proportions separated by an invasion barrier. As for the case with no observation costs, results are very insensitive to r. As the cost of observation increases, the range of values of p for which player 1 rejects player 2 without observation increases. Thus, the position of the invasion barrier increases as costs increase. The evolutionarily stable proportion of trustworthy individuals decreases as costs increase.

(e) *Signalling by player 2*

In the model analysed above, player 1 observes a signal from player 2, but a player 2 of a given type has had no control over the signal sent. We now analyse a variant of the model in which player 2 has influence over a costly signal. As in the basic trust model player 1 cannot control the quality of the signal received from player 2. Player 1 just observes the signal of player 2 at no cost and then decides whether or not to trust player 2. Player 2 decides the mean of the signal that is sent. The actual signal sent by player 2 is drawn from a normal distribution with fixed variance $\sigma^2 = 1$. The signalling cost paid by player 2 depends on type and the mean signal chosen as follows. If an untrustworthy individual chooses mean $\bar{\mu}$ (where $\bar{\mu} \geqslant -1$) then this individual pays cost

$$\bar{K}(\bar{\mu}) = k(\bar{\mu} + 1)^2. \tag{3.8}$$

If a trustworthy individual chooses mean μ (where $\mu \geqslant 1$) then this individual pays cost

$$K(\mu) = k(\mu - 1)^2. \tag{3.9}$$

Thus, we are assuming that untrustworthy individuals have a 'baseline signal' whose mean is –1 and trustworthy individuals have a baseline signal with mean +1. In both cases it costs to increase the mean signal from its baseline value, with cost increasing quadratically with the deviation. Here, k is a parameter determining the absolute magnitude of the cost paid.

We solve for the evolutionarily stable population composition as follows. For a given proportion of trustworthy individuals, p, the three variables $x_c(p)$, $\bar{\mu}(p)$ and $\mu(p)$ satisfy the following consistency conditions: (i) $x_c(p)$ is the optimal critical acceptance threshold for player 1 given that untrustworthy and trustworthy individuals choose mean signals $\bar{\mu}(p)$ and $\mu(p)$, respectively; (ii) $\bar{\mu}(p)$ is the optimal signal mean for untrustworthy individuals given player 1 uses threshold $x_c(p)$; and (iii) $\mu(p)$ is the optimal signal mean for trustworthy individuals given player 1 uses threshold $x_c(p)$. From these variables the pay-offs to each type of player 2 can be evaluated. The evolutionarily stable proportion of trustworthy individuals can then be determined as for the basic model.

Figure 5 illustrates a result that is robust provided the cost parameter k is not too small. When p is small, trustworthy individuals raise their mean signal (above their baseline) by much more than untrustworthy individuals raise their signal (above their baseline). This result can be understood as follows. When p is small the acceptance threshold of player 1 is high. Raising $\bar{\mu}$ by a small amount from its baseline value of –1 then makes little difference to the probability that an untrustworthy individual is trusted, but raising μ from its baseline value of +1 has an appreciable effect. When p is large acceptance thresholds

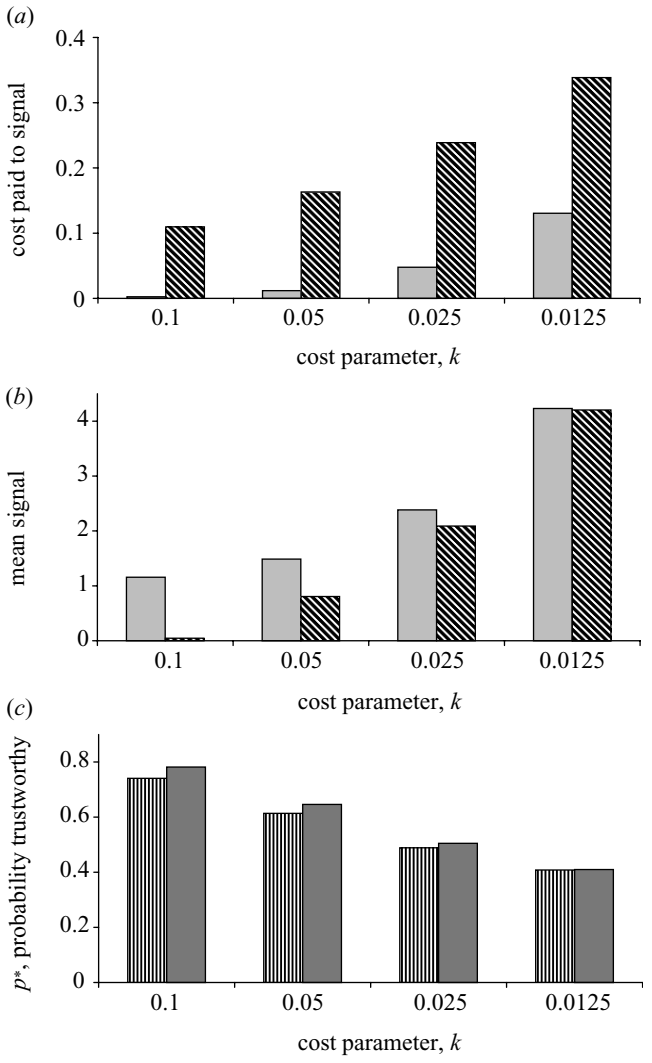

Figure 6. The effect of the cost of changing the mean signal in the model in which player 2 can adjust the mean of the signal it gives. Results shown are at evolutionary stability. (*a*) The cost paid by each type of player 2: shaded bars, trustworthy; hatched bars, untrustworthy. (*b*) The mean signal of each type of player 2: shaded bars, trustworthy; hatched bars, untrustworthy. (*c*) The evolutionarily stable proportion of trustworthy individuals, p^* (lined bars), and the probability that an individual that is trusted by player 1 is trustworthy (shaded bars). Parameters are $s = 0.3$; $r = 0.75$, $\sigma = 1$.

are low and the reverse is true, but typically the discrepancy between the types is more marked. At evolutionary stability the signalling cost paid by untrustworthy individuals is much larger than the signalling cost paid by trustworthy individuals. Note that because both types do equally well at stability, this means that untrustworthy individuals are doing better at getting the game pay-offs.

As the cost parameter k is reduced, individuals of both types increase their mean signal. However, at equilibrium untrustworthy individuals pay a greater signalling cost than trustworthy individuals (figure 6*a*). As a consequence signal means are close at equilibrium (figure 6*b*). Thus, when k is small player 1 has difficulty in differentiating between the two types (figure 6*c*). The equilibrium value of p is then close to the critical proportion for acceptance when there is no observation of type, s/r (equation (3.3)).

4. DISCUSSION

In this paper we have been concerned with whether animals can make reliable threats and promises. When a particular course of action is in an animal's best interests, then it is obviously believable that the animal will adopt this behaviour. In the context of contests between animals, Enquist (1985) showed that at the ESS, animals could give reliable information about their future behaviour. In his examples, at the ESS each animal adopts the best action given its current circumstances, i.e. the equilibrium is sub-game perfect. In our example based on desertion, we described how introducing energy reserves as a state variable made it possible for females to create a credible threat to desert if the male deserted. The female's threat is based on the fact that her best action if deserted when she has low reserves is to desert. In other words, her strategy is sub-game perfect. The female's manipulation of her reserves is an example of the sort of commitment tactics considered by Schelling (1960):

> The essence of these tactics is some voluntary but irreversible sacrifice of freedom of choice. They rest on the paradox that the power to constrain an adversary may depend on the power to bind oneself.…
> (Schelling1960, p. 22)

Sub-game perfection requires an animal to behave optimally in any circumstances in which it finds itself. This requirement does not strike us as realistic. One challenge to sub-game perfection is based on the idea that animals follow relatively simple rules that perform well in most of the circumstances that the animal encounters. Rather than having a specific response for every contingency, various classes of outcome may be treated in a similar way (e.g. Enquist *et al.* 2002), perhaps by inducing a particular state of the animal. These states can be regarded as emotional states (cf. Trivers 1971; Leimar 1997). In this view, emotions are a component of the rules that determine consistent patterns of behaviour. The behaviour of an animal is characterized by its states and the rules that determine both how states change (as a function of current state and the environment, including other animals) and how state determines behaviour. Individual members of a species may have broadly similar emotional states but may differ in the associated rules. As a result, individuals will differ in their typical patterns of behaviour in a range of contexts. We refer to such a consistent pattern as a personality. Thus, an emotion is a relatively short-lived state that gives some information about behaviour in the immediate future, whereas personality is a fundamental characteristic that can provide information about behaviour in a variety of contexts. Darwin (1872) drew attention to reliable associations between intentions and cues: 'when a dog approaches a strange dog or man in a savage frame of mind he walks upright and very stiffly; his head slightly raised, or not much lowered; the tail is held erect and quite rigid; the hairs bristle, especially along the neck and back; the pricked ears are directed forward, and the eyes have a fixed stare… These actions, as will hereafter be explained, follow from the dog's intention to attack his enemy, and are thus to a large extent intelligible.' (pp. 55–56). In his discussion of lying, Ekman (1985) argues that humans may erroneously use cues about emotion as indicators of personality:

> the stereotype that a thin-lipped person is cruel is based on the accurate clue that lips do narrow in anger. The error is in utilizing a sign of temporary emotional state as the basis for judging a personality *trait*. (p. 26)

Because they provide information about future behaviour, it may be advantageous for an animal to obtain information about the emotional state or the personality of an animal before engaging in an important interaction. It may also be advantageous for one animal to mislead another about its emotional state or personality. Frank (1987–1989) uses the idea of reliable indicators to argue that cooperation behaviour can be maintained in a population despite the temptation to cheat. Although Frank allows for the possibility that acquiring a reputation for behaving in a particular way may be important, he shows that cooperation is possible when there is a single interaction and hence no effect of reputation. Like Frank (1987) and Guth & Kliemt (2000), we consider this case. There are two types of individual in the population; trustworthy and untrustworthy. When two individuals meet, one player can observe the other and on the basis of the observation decides whether to trust the other player or reject it. If trusted, a trustworthy player cooperates, whereas an untrustworthy player defects. If an individual is trusted, it gets a higher pay-off from defecting than from cooperating, so the trustworthy individuals do not make the best choice when they are trusted. The fundamental assumption is that they behave in a consistently trustworthy way despite the temptation to defect. At the equilibrium, both types do equally well. Once accepted, untrustworthy individuals do better than trustworthy individuals, but trustworthy individuals have a better chance of being accepted. This arises from the fact that observations provide information about type. Guth & Kliemt assume that observations can only take one of two values. By contrast, we assume that observations are normally distributed about a mean that depends on the observed individual's type. Although we talk about one individual making an observation that provides information about another individual's type, the process of obtaining information may actually involve a series of interactions between the two individuals. We make no attempt to model this interaction: we use the single observation with a normally distributed outcome as a simple and convenient characterization of the observation process.

The observation process captures the idea that it is possible to obtain information about the type (or personality) of an individual. Given that one player is trying to determine the type of the other player and prefers to interact with trustworthy individuals, there will be pressure on untrustworthy individuals to resemble trustworthy individuals. Our analysis goes beyond that of Guth & Kliemt (2000) in that we consider the evolution of the appearance of each type of individual. Each type can choose the 'signal' that is sent to the observing individual, but the cost of the signal increases as the signal becomes more unlike the type's baseline value. This evolution can be thought of as an arms race (Dawkins & Krebs 1979) between the two types, in which the untrustworthy individuals try to mimic the trustworthy individuals. Our general conclusion is that at equilibrium, untrustworthy individuals pay a greater signalling cost than trustworthy individuals. We emphasize that our results on this topic are preliminary, and that a complete analysis should allow for the evolution of the cost devoted to observation.

We thank Anthony Arak, Sasha Dall, Magnus Enquist and Alex Kacelnik for their comments on a previous version of this manuscript.

APPENDIX A: CRITICAL ACCEPTANCE THRESHOLDS

Let the random variable X be observed by player 1. This random variable has probability density function $f(x)$ when player 2 is trustworthy and density $\bar{f}(x)$ when player 2 is not trustworthy. We assume that that the ratio $f(x)/\bar{f}(x)$ is a strictly increasing function of x. Let $W(x)$ be the expected pay-off to player 1 if this player accepts player 2 if and only if the observation of this player exceeds x. Then

$$W(x) = p[s\,P(X \leqslant s) + r\,P(X > x)] + (1-p)\;[s\,\bar{P}(X \leqslant s) + 0.\bar{P}(X > x)], \qquad \text{(A 1)}$$

where P denotes the probability under f and $\bar{P}$ denotes the probability under $\bar{f}$. Differentiating with respect to x we deduce that $W(x)$ is maximized when $x = x_c(p)$ where

$$\frac{f(x_c(p))}{\bar{f}(x_c(p))} = \left(\frac{(1-p)}{p}\right)\left(\frac{s}{r-s}\right). \qquad \text{(A 2)}$$

We now take f and $\bar{f}$ to be the density functions of normally distributed random variables with means μ and $\bar{\mu}$, respectively. In each case the variance is σ^2. Then by equation (A 2) we have

$$x_c(p) = \frac{(\mu + \bar{\mu})}{2} + \frac{\alpha\,\sigma^2}{(\mu - \bar{\mu})}, \qquad \text{(A 3)}$$

where

$$\alpha = \log\left(\frac{1-p}{p}\right) + \log\left(\frac{s}{r-s}\right). \qquad \text{(A 4)}$$

Equation (3.5) is obtained by setting $\bar{\mu} = -1$, $\mu = 1$ and $\sigma^2 = 1$.

APPENDIX B: GENERAL PROPERTIES AT EVOLUTIONARY STABILITY

We consider the trust model with fixed signal by each type of player 2. The signal can have any distribution and does not have to be normal. The analysis applies whether or not there is an observation cost to player 1.

Consider a population at evolutionary stability. Let p^* be the proportion of trustworthy individuals in the population. Let q be the probability that a trustworthy individual is trusted. Then the pay-off to a trustworthy individual is

$$(1-q)\,s + q\,r. \qquad \text{(B 1)}$$

Similarly the pay-off to an untrustworthy individual is

$$(1-\bar{q})\,s + \bar{q}, \qquad \text{(B 2)}$$

where $\bar{q}$ is the probability that an untrustworthy individual is trusted. Because the pay-offs to each type of individual are equal at evolutionary stability we have

$$q\,(r-s) = \bar{q}\,(1-s), \qquad \text{(B 3)}$$

which is equation (3.6) of the main text.

By Bayes theorem

$$P(\text{trustworthy}|\text{trusted}) = \frac{p^*q}{p^*q + (1 - p^*)\,\bar{q}}\ . \qquad \text{(B 4)}$$

By equation (3.4) this probability exceeds s/r. Thus,

$$p^*q\,r > s\,(p^*q + (1 - p^*)\bar{q}). \qquad \text{(B 5)}$$

Substituting for q from equation (B 3) then gives equation (3.7) of the main text.

APPENDIX C: OBSERVATION COSTS

Consider the model in which player 1 pays an observation cost to observe player 2. We assume that if the cost paid is c then the standard deviation of the observation is

$$\sigma = \frac{1}{\left(\frac{c}{c_0}\right)^{\frac{1}{8}} + \left(\frac{c}{c_0}\right)^{\frac{1}{2}}}, \qquad \text{(C 1)}$$

where c_0 is a positive parameter.

REFERENCES

Barta, Z., Houston, A. I., McNamara, J. M. & Szekely, T. 2002 Sexual conflict about parental care: the role of reserves. *Am. Nat.* **159**, 687–705.

Bradbury, J. W. & Vehrencamp, S. L. 1998 *Principles of animal communication*. Sunderland, MA: Sinauer.

Brams, S. 1983 *Superior beings*. New York: Springer.

Budaev, S. V., Zworykin, D. D. & Mochek, A. D. 1999 Individual differences in parental care and behaviour profile in the convict cichlid: a correlation study. *Anim. Behav.* **58**, 195–202.

Darwin, C. 1872 *The expression of the emotions in man and animals* (ed. P. Ekman, 1998). Oxford University Press.

Dawkins, R. & Krebs, J. R. 1979 Arms races between and within species. *Proc. R. Soc. Lond.* B **205**, 489–511.

Ekman, P. 1985 *Telling lies*. New York: Norton.

Elster, J. 1984 *Ulysses and the sirens*. Cambridge University Press.

Elster, J. 2000 *Ulysses unbound*. Cambridge University Press.

Enquist, M. 1985 Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Anim. Behav.* **33**, 1152–1161.

Enquist, M., Arak, A., Ghirlanda, S. & Wachtmeister, C.-A. 2002 Spectacular phenomena and limits to rationality in genetic and cultural evolution. *Phil. Trans. R. Soc. Lond.* B **357**. (In this issue.) (DOI 10.1098/rstb.2002.1067.)

Frank, R. H. 1987 If *Homo economicus* could choose his own utility function, would he want one with a conscience? *Am. Econ. Rev.* **77**, 593–604.

Frank, R. H. 1988 *Passions within reason*. New York: Norton.

Frank, R. H. 1989 Honesty as an evolutionarily stable strategy. *Behav. Brain Sci* **12**, 705–706.

Fudenberg, D. & Tirole, J. 1991 *Game theory*. Cambridge MA: MIT Press.

Guth, W. & Kliemt, H. 2000 Evolutionarily stable cooperative commitments. *Theory and Decision* **49**, 197–221.

Heinsohn, R. & Packer, C. 1995 Complex cooperative strategies in group-territorial African lions. *Science* **269**, 1260–1262.

Houston, A. I. & McNamara, J. M. 1999 *Models of adaptive behaviour*. Cambridge University Press.

Houston, A. I., Kacelnik, A. & McNamara, J. M. 1982 Some learning rules for acquiring information. In *Functional ontogeny* (ed. D. J. McFarland), pp. 140–191. London: Pitman.

Hurd, P. L. & Enquist, M. 1998 Conventional signaling in aggressive interactions: the importance of temporal structure. *J. Theor. Biol.* **192**, 197–211.

Johnstone, R. A. 1997 The evolution of animal signals. In *Behavioural ecology*, 4th edn (ed. J. R. Krebs & N. B. Davies), pp. 155–178. Oxford: Blackwell Science.

Johnstone, R. A. 1998 Game theory and communication. In *Game theory and animal behavior* (ed. L. A. Dugatkin & H. K. Reeve), pp. 94–117. New York: Oxford University Press.

Krebs, J. R. & Dawkins, R. 1984 Animal signals: mind-reading and manipulation. In *Behavioural ecology*, 2nd edn (ed. J. R. Krebs & N. B. Davies), pp. 380–402. Oxford: Blackwell Scientific.

Leimar, O. 1997 Repeated games: a state-space approach. *J. Theor. Biol.* **184**, 471–498.

McNamara, J. M. 1996 Risk-prone behaviour under rules which have evolved in a changing environment. *Am. Zool.* **36**, 484–495.

McNamara, J. M. & Houston, A. I. 1980 The application of statistical decision theory to animal behaviour. *J. Theor. Biol.* **85**, 673–690.

McNamara, J. M., Gasson, C. E. & Houston, A. I. 1999 Incorporating rules for responding into evolutionary games. *Nature* **401**, 368–371.

Maynard Smith, J. 1979 Game theory and the evolution of behaviour. *Proc. R. Soc. Lond.* B **205**, 475–488.

Maynard Smith, J. 1982 *Evolution and the theory of games*. Cambridge University Press.

Samuelson, L. 2001 Introduction to the evolution of preferences. *J. Econ. Theory* **97**, 225–230.

Schelling, T. C. 1960 *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Selten, R. 1980 A note on evolutionary stable strategies in asymmetric animal conflicts. *J. Theor. Biol.* **84**, 93–101.

Selten, R. 1983 Evolutionary stability in extensive two-person games. *Math Soc. Sci.* **5**, 269–363.

Trivers, R. L. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57.

Wilson, D. S. 1998 Adaptive individual differences within single populations. *Phil. Trans. R. Soc. Lond.* B **353**, 199–205. (DOI 10.1098/rstb.1998.0202.)

Wilson, D. S., Clark, A. B., Coleman, K. & Dearstyne, T. 1994 Shyness and boldness in humans and other animals. *Trends Ecol. Evol.* **9**, 442–446.

GLOSSARY

ESS: evolutionarily stable strategy